BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e26539
doi: 10.3897/biss.2.26539

OPEN ACCESS

Conference Abstract

Kurator: Tools for Improving Fitness for Use of Biodiversity Data.

Paul J. Morris‡, James Hanken‡, David B. Lowery‡, Bertram Ludäscher§, James Macklin|, Timothy McPhillips§, John Wieczorek¶, Qian Zhang§

‡ Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States of America
§ University of Illinois Urbana-Champaign, Champaign, United States of America
| Agriculture and Agri-Food Canada, Ottawa, Canada
¶ Museum of Vertebrate Zoology, University of California, Berkeley, United States of America

Corresponding author: Paul J. Morris (mole@morris.net)

Received: 09 May 2018 | Published: 15 Jun 2018

Citation: Morris P, Hanken J, Lowery D, Ludäscher B, Macklin J, McPhillips T, Wieczorek J, Zhang Q (2018) Kurator: Tools for Improving Fitness for Use of Biodiversity Data. Biodiversity Information Science and Standards 2: e26539. https://doi.org/10.3897/biss.2.26539

Abstract

As curators of biodiversity data in natural science collections, we are deeply concerned with data quality, but quality is an elusive concept. An effective way to think about data quality is in terms of fitness for use (Veiga 2016). To use data to manage physical collections, the data must be able to accurately answer questions such as what objects are in the collections, where are they and where are they from. Some research uses aggregate data across collections, which involves exchange of data using standard vocabularies. Some research uses require accurate georeferences, collecting dates, and current identifications. It is well understood that the costs of data capture and data quality improvement increase with increasing time from the original observation. These factors point towards two engineering principles for software that is intended to maintain or enhance data quality: build small modular data quality tests that can be easily assembled in suites to assess the fitness of use of data for some particular need; and produce tools that can be applied by users with a wide range of technical skill levels at different points in the data life cycle.

In the Kurator project, we have produced code (e.g. Wieczorek et al. 2017, Morris 2016) which consists of small modules that can be incorporated into data management

processes as small libraries that address particular data quality tests. These modules can be combined into customizable data quality scripts, which can be run on single computers or scalable architecture and can be incorporated into other software, run as command line programs, or run as suites of canned workflows through a web interface. Kurator modules can be integrated into early stage data capture applications, run to help prepare data for aggregation by matching it to standard vocabularies, be run for quality control or quality assurance on data sets, and can report on data quality in terms of a fitness-for-use framework (Veiga et al. 2017). One of our goals is simple tests usable by anyone anywhere.

Keywords

data quality, workflows, biodiversity informatics, natural science collections

Presenting author

Paul J. Morris

Presented at

SPNHC 2018

Funding program

US NSF:ABI (Advances in Biological Informatics)

Grant title

Collaborative Research: ABI Development: Kurator: A Provenance-enabled Workflow Platform and Toolkit to Curate Biodiversity Data 1356438 and 1356751

References

- Morris P (2016) FilteredPush/event_date_qc: Release version 1.0.2 (Version v1.0.2). Zenodo https://doi.org/10.5281/zenodo.210631
- Veiga AK (2016) *A conceptual framework on biodiversity data quality*. Tese de Doutorado, Escola Politécnica, Universidade de São Paulo, São Paulo, Recuperado em 2018-03-26 pp. URL: http://www.teses.usp.br/teses/disponiveis/3/3141/tde-17032017-085248/

- Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D, Robertson TJ (2017) A conceptual framework for quality assessment and management of biodiversity data. PLOS ONE 12 (6): e0178731. https://doi.org/10.1371/journal.pone.0178731
- Wieczorek J, McPhillips T, Lowery D, Morris P, Zhang Q, ElSkunkito (2017) kurator-org/kurator-validation: Kurator-validation version v1.0.1 (Version v1.0.1). Zenodo https://doi.org/10.5281/zenodo.1068330